COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Fall 2026.

Lecture 10

Midterm Logistics

- Project 2 due Friday 11:59pm
- Midterm 1 is next Monday in class.
- Closed book, no calculator (will be designed so neither are needed).
- Study guide is posted on the course webpage (under the midterm row in the schedule tab).
- Past midterms are posted in Canvas. Note that some cover more material than we have seen – e.g., we have not covered locality sensitive hashing, the Johnson-Lindenstrauss lemma or low-distortion embeddings.
- I am holding office hours **Friday 11:30am-12:30pm**.

Studying

Suggested Studying Approach:

- Review the study guide to get a sense of what you need to know, and then mostly focus on doing practice questions from the past midterms and the study guide.
- Review slides only as needed.
- Make sure you understand the solutions to all problem set and quiz questions.
- If you get nervous during exams, do some practice exams under time constraints with no material in front of you.
- Note that many of the past exams were 2 hours long, but designed to be completed in 90 minutes. This exam will be a bit shorter.

Midterm Format

Rough Outline: (subject to change)

- Question 1: 4-5 True/False or ALWAY/SOMETIMES/NEVER questions.
- Question 2: 4-5 short answers, like quiz questions.
- Question 3: 4-5 part question on analyzing an algorithm. Similar in style to but easier than a homework question.
- Question 4: More challenging 4-5 part question on analyzing an algorithm.

A few extra credit points available on Q3 and Q4.

Questions

Content or Format Questions?

- streaming algos, where we can use certain hash functions

- space complexity for streaming

2 level hashing:

$\mathbb{E}[S_i^2]$ (size i^{th} backup table)

$\mathbb{E}\left[\left(\sum S_{ij}\right)^2\right] = \mathbb{E}\sum_k\sum_j S_{ij}S_{ik} = \sum_k\sum_j \mathbb{E}[S_{ij}S_{ik}]$ ← used pairwise ind.

$\sum_i \mathbb{E}\,S_i^2 = \sum_i\sum_k\sum_j \mathbb{E}S_{ij}S_{ik} = \sum_k\sum_j\sum_i \mathbb{E}S_{ik}S_{jk}$

? $\Pr[h(x_j)=h(x_j)]$

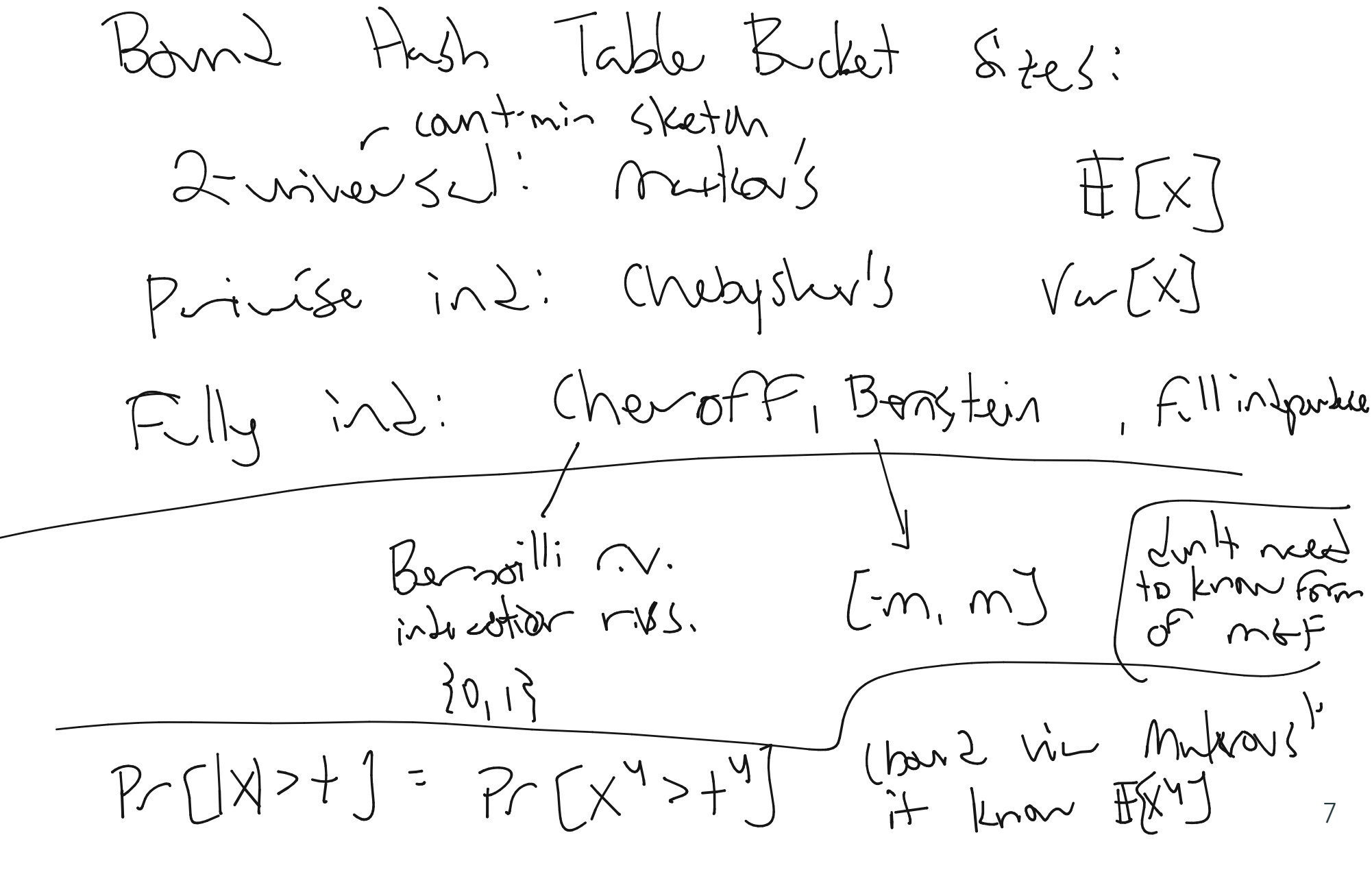
Questions

$$Pr[S_i > t] \leq \frac{1}{10}$$

how big do I need to set t if I use:

- Markov
- Chebyshev
- Chernoff

Questions


Bound Hash Table Bucket Sizes:
count-min sketch
2-universal: Markov's
𝔼[X]
Pairwise ind: Chebyshev's
Var[X]
Fully ind: Chernoff, Bernstein, fill in paradise
Bernoulli r.v.
indicator rvs.
{0,1}
[-m, m]
don't need to know form of MGF
Pr[|X|>t] = Pr[X^4>t^4]
(bound via Markov's)
it know 𝔼[X^4]

Questions

- pairwise independence suffices for linearity of variance

hash n items to n-slot hash table

S_i = # items in slot i

$\mathbb{E} S_i = 1$ (fully random, pairwise ind. 2 universal)

$\mathrm{Var}[S_i] = \mathrm{Var}\left[\sum_{j=1}^{n} S_{ij}\right]$ S_{ij} = indicator that item j hashes to slot i

limits of variance

$= \sum_{j=1}^{n} \mathrm{Var}[S_{ij}]$ (fully random, pairwise ind)

≤ 1 $\quad$ 1 w.p. $\frac{1}{n}$

0 w.p. $1-\frac{1}{n}$

not for 2-universal

$\mathrm{Var}[S_i] = \mathbb{E}S_{ij}^2 - [\mathbb{E}S_{ij}]^2 \approx \frac{1}{n}$

$\frac{1}{n} - \frac{1}{n^2}$

Questions

2 universal function: (low collision prob)

$$\Pr_{x \neq y}(h(x) = h(y)) \leq \frac{1}{n}$$

- count-min sketch
- 2 level hashing → in class we used pairwise ind. can generalize to use 2 universal $\mathbb{E}$

Pairwise independent: $x \neq y$

$$\Pr(h(x) = i \cap h(y) = j) = \frac{1}{n^2} = \Pr(h(x) = i) \cdot \Pr(h(y) = j)$$

$$\Pr(h(x) = i) \cap h(y) = j \cap h(z) = k) \neq \frac{1}{n^3}$$

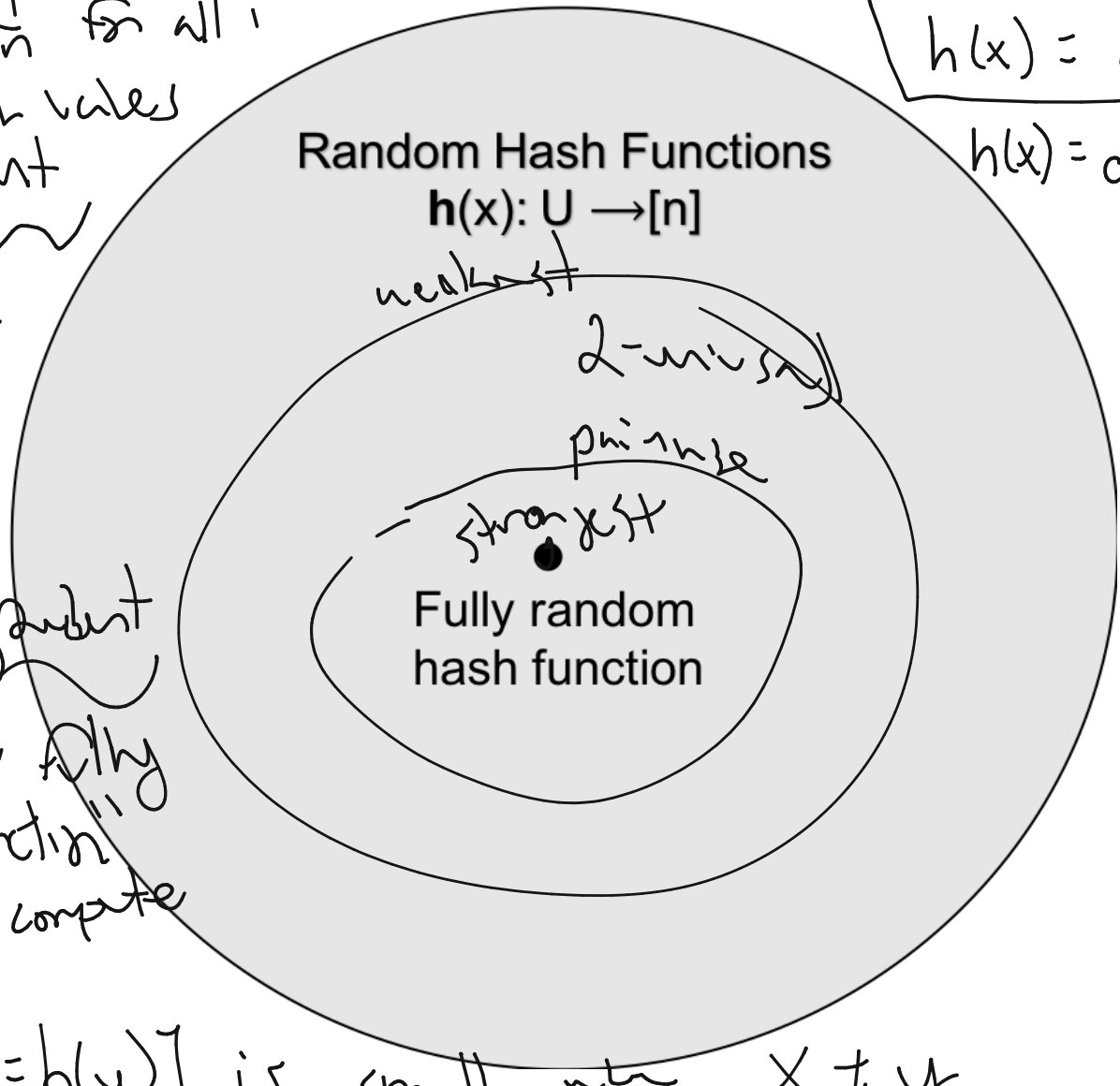
Random Hash Functions

$Pr[h(x) = i] = \frac{1}{n}$ for all i and all hash values are independent

↳ to expensive

$h(x) = ax + b \bmod n$

$h(x) = ax^2 + bx + c \bmod c$



2-universal

p-wise independent

"approximate fully random function"

- difficult to compute

$Pr(h(x) = h(y)]$ is small when $x \neq y$

Concentration Bounds

Concentration Bound Requirements

Markov's	Chebyshev's	Chernoff	Bernstein

Sampling For Mean Estimation

Say I have n numbers $x_1, \ldots, x_n$ all lying in $[-M, M]$ with mean $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$. How can I estimate μ without reading all the numbers?

Johnson-Lindenstrauss Proof Outline

MinHashing For Distinct elts:

Basic algo:

$s_1 \ldots \ldots s_t = \min_{i=1..n} h_j(x_i)$

$h_1 \ldots h_t$

space to compute $O(t)$

$s = \frac{1}{t} \sum_{j=1}^{t} s_j$

$\mathbb{E}S = \frac{1}{d+1}$ $\quad Var(S) \leq \frac{1}{(d+1)^2 t}$

if $t \geq \frac{1}{\varepsilon^2 \delta}$

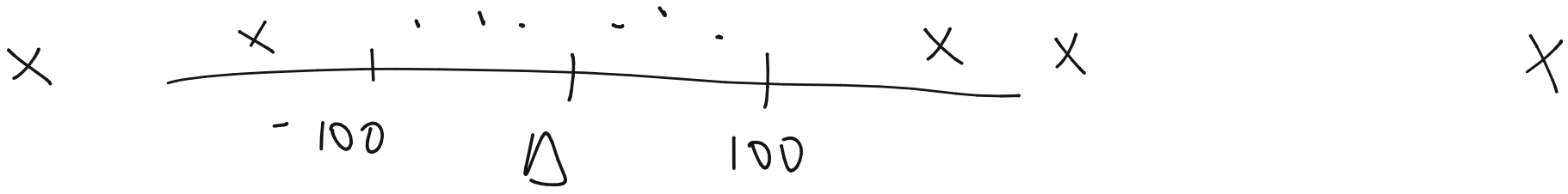
by Chebyshev

$\Pr[|S - \mathbb{E}S| \geq \varepsilon \mathbb{E}[S]] \leq \delta$

$O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$ median trick $O\left(\frac{1}{\varepsilon^2 \delta}\right)$

Median Trick

3. Consider an algorithm $\mathcal{A}$ running in time $T(\mathcal{A})$, that with probability .6 outputs an estimate of the number of triangles in an input graph up to error ± 100, and with probability .4 outputs some bad estimate with worse error. Describe an algorithm that outputs an estimate of the number of triangles in an input graph up to error ± 100 with probability $\geq .99$ and runs in time $O(T(\mathcal{A}))$.



run A t times independently get estimates $\Delta_1 \ldots \Delta_t$

X = # times that apx in $\Delta \pm 100$

if $X > \frac{t}{2}$ then median $(\Delta_1, \ldots \Delta_t) \in \Delta \pm 100$

$\Pr\left(X > \frac{t}{2}\right) > .99$ then I'm done

concentration bound

The Chernoff bound states that for independent random variables $X_1, \ldots, X_n$ taking values in $\{0, 1\}$, letting $\mu = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]$, for any $\delta > 0$,

$\Pr\left(\left|\sum_{i=1}^{n} X_i - \mu\right| > \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2\mu}{2+\delta}\right)$.

Median Trick

output median($D_1, \ldots, D_t$) — running time $t \cdot T(A)$ + $O(T(A))$

3. Consider an algorithm $\mathcal{A}$ running in time $T(\mathcal{A})$, that with probability .6 outputs an estimate of the number of triangles in an input graph up to error ± 100, and with probability .4 outputs some bad estimate with worse error. Describe an algorithm that outputs an estimate of the number of triangles in an input graph up to error ± 100 with probability $\geq .99$ and runs in time $O(T(\mathcal{A}))$.

$\delta \leq \frac{1}{3}$

$t = \frac{25}{\delta}$

$\Pr\left(X \geq \frac{t}{2}\right) \geq .99$

X = # of "good" estimates in t total tries

$X_i = 1$ if trial i is good, 0 otherwise

(Chernoff: $\Pr\left(X < \frac{t}{2}\right) \leq .01$ for $t = c\log(1/\delta)$)

$\mathbb{E}[X] = .6t$

$\mathrm{Var}(X) = \sum_{i=1}^{t} \mathrm{Var}(X_i) \leq t(.6 - .6^2) \leq .25t$

$\Pr(X - \mathbb{E}X < -.1t)$

$\leq \Pr(|X - \mathbb{E}X| \geq .1t)$ by Chebyshev $\leq \frac{.25t}{(.1t)^2} = \frac{.25}{.01t} = \frac{25}{t} \leq .01$ for $t = 2500$

which bounds? ~~Markov~~, Chebyshev, Chernoff, ~~Bernstein~~

The Chernoff bound states that for independent random variables $X_1, \ldots, X_n$ taking values in $\{0, 1\}$, letting $\mu = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]$, for any $\delta > 0$, $\Pr\left(\left|\sum_{i=1}^{n} X_i - \mu\right| > \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2\mu}{2+\delta}\right)$.

Y = # bad estimates

$\mathbb{E}Y = .4t$

$\Pr(Y \geq \frac{t}{2}) \leq .01$

Example Problems

2. Assume there are 1000 registered users on your site $u_1, \ldots, u_{1000}$, and in a given day, each user visits the site with some probability p_i. The event that any user visits the site is independent of what the other users do. Assume that $\sum_{i=1}^{1000} p_i = 500$.

   (a) Let $\mathbf{X}$ be the number of users that visit the site on the given day. What is $\mathbb{E}[\mathbf{X}]$.

   (b) Apply a Chernoff bound to show that $\Pr[\mathbf{X} \geq 600] \leq .01$.

   (c) Apply Markov's inequality and Chebyshev's inequality to bound the same probability. How do they compare?

The Chernoff bound states that for independent random variables $X_1, \ldots, X_n$ taking values in $\{0, 1\}$, letting $\mu = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]$, for any $\delta > 0$,
$\Pr\left(\left|\sum_{i=1}^{n} X_i - \mu\right| > \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2\mu}{2+\delta}\right)$.

Example Problems

2. Assume there are 1000 registered users on your site $u_1, \ldots, u_{1000}$, and in a given day, each user visits the site with some probability p_i. The event that any user visits the site is independent of what the other users do. Assume that $\sum_{i=1}^{1000} p_i = 500$.

   (a) Let $\mathbf{X}$ be the number of users that visit the site on the given day. What is $\mathbb{E}[\mathbf{X}]$.

   (b) Apply a Chernoff bound to show that $\Pr[\mathbf{X} \geq 600] \leq .01$.

   (c) Apply Markov's inequality and Chebyshev's inequality to bound the same probability. How do they compare?

The Chernoff bound states that for independent random variables $X_1, \ldots, X_n$ taking values in $\{0, 1\}$, letting $\mu = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]$, for any $\delta > 0$,
$\Pr\left(\left|\sum_{i=1}^{n} X_i - \mu\right| > \delta\mu\right) \leq 2\exp\left(-\frac{\delta^2\mu}{2+\delta}\right)$.

Example Problems

ALWAYS, SOMETIMES, or NEVER:

$O(t \cdot \log\log d + \log d)$

2. $\Pr[\max(X_1, \ldots X_n) \geq t] \leq \sum_{i=1}^{n} \Pr[X_i \geq t]$ for any random variables $X_1, \ldots, X_n$.

(c) $\Pr[\mathbf{X} = s \cap \mathbf{Y} = t] = \Pr[\mathbf{X} = s] \cdot \Pr[\mathbf{Y} = t]$.

loglog, hyperloglog

min-hash algo $h: U \to [0,1]$

- use discrete hash functions
- work better in practice

$X_1 \to [0\,1\,1\,0\,0\,1\,1]$

$X_2 \to [0\,0\,1\,0\,0\,0\,1]$

$\vdots$

$X_n = [0\,1 \ldots 0\,0]$ — max # trailing 0s

[max trailing 0s] = log(d)

store m [0, log(d)]

m tasks (log log d) bit to store

Example Problems

ALWAYS, SOMETIMES, or NEVER:

- $\mathsf{Var}(X + Y) < \mathsf{Var}(X) + \mathsf{Var}(Y)$.

- $\mathbb{E}[1/X] = 1/\mathbb{E}[X]$.